MetaMap Migration to Lexical Tools Java APIs – ASCII Issues

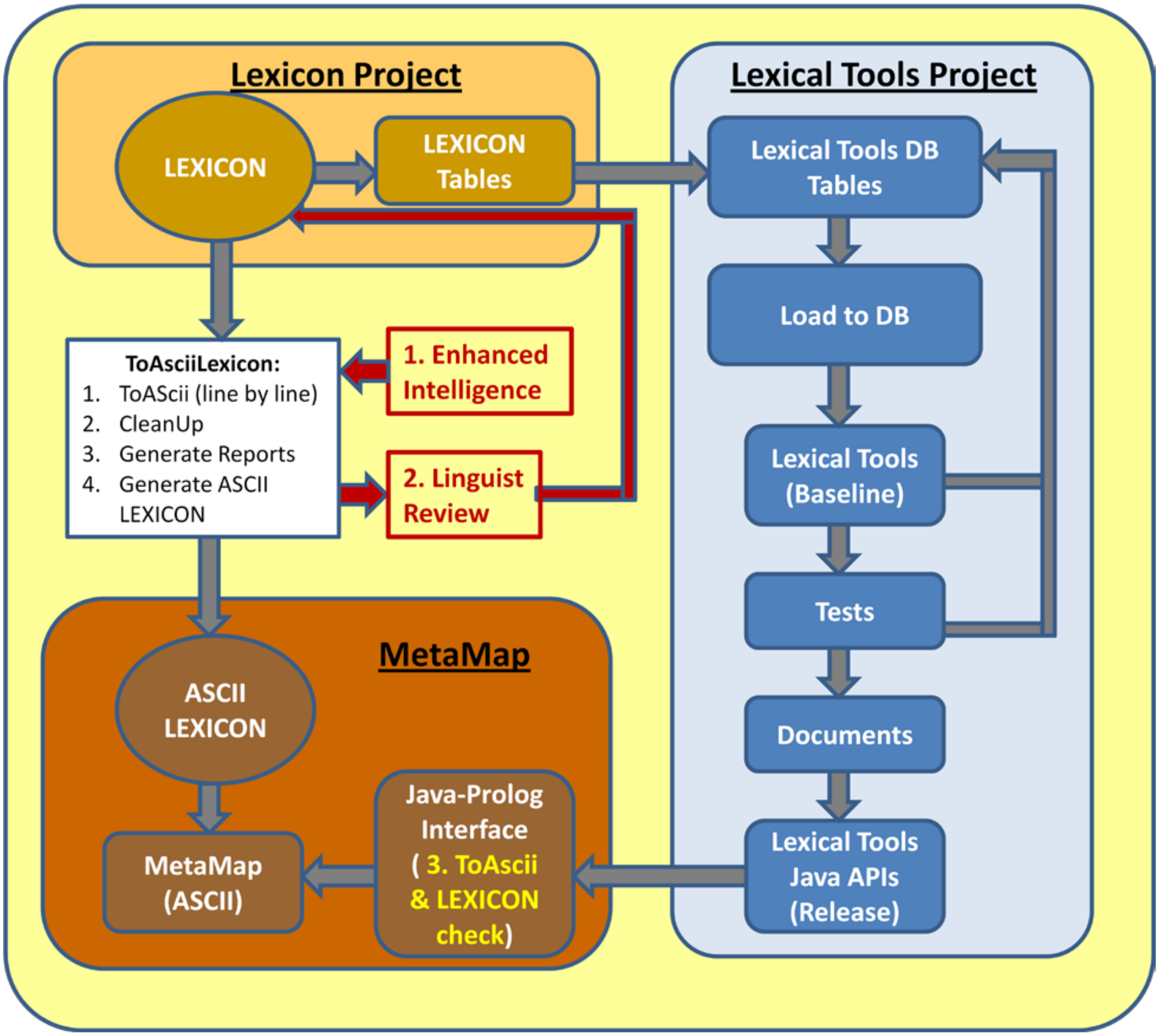
I. ASCII LEXICON

Lexical System Group generates a pure ASCII LEXICON from The SPECIALIST LEXICON to support MetaMap project annually. A program, ToAsciiLexicon, was developed and used for this purpose. It does 1). Convert all contents to pure ASCII; 2). Remove duplications after ASCII conversion, such as spelling variants, abbreviations, acronyms, etc.; 3). Generate reports on all ASCII conversion; 4) Generate pure ASCII LEXICON. In 2010 release, there are 432,822 records in both LEXICON and ASCII LEXICON. In other words, ToAsciiLexicon program does not remove any lexical record. A statistics data of ASCII conversion are shown in the table below:

Fields	**ASCII conversion**	**Examples**
base=	299 ⇨ ***29 not-Lex:*** 26 + 3 (diff LexRec) ⇨ 26 test: 23 + 3 ASCII spVars	**E0586236**: base=styrenatíon ⇨ Styrenation (not in LEXICON)
spelling_variant=	3955 ⇨ ***284 not-Lex:*** 260 + 24 (AE, ..) ⇨ 387 test: 260 + 7 (R) + 120 (C)	**E0041164**: spelling_variant=µm ⇨ mum (in LEXICON, but different record)
variants=irreg|…	67 ⇨ 66 Duplicated ⇨ ***1 not-Lex (folumæ)***	**E0028609**: variants=irreg|formula|formulæ| • formulae (not in LEXICON, but OK in Unicode DB)
acronym_of=	26 ⇨ ***6 not-Lex***	**E0501790**: acronym_of=Sjögren's syndrome A|E0632187 • Sjogren's syndrome A (in LEXICON)
abbreviation_of=	5 ⇨ ***1 not-Lex***	**E0632532**: abbreviation_of=interferon-stimulated gene factor 3α|E0632531 • interferon-stimulated gene factor 3alpha (not in LEXICON)
compl=	2 ⇨ ***2 not-Lex***	**E0027351**:compl=pphr(of,np|Türck|) • Turck (in LEXICON), but Türck is not in Lexicon
nominalization=	3 ⇨ ***2 not-Lex***	**E0604056**: nominalization=nonnaïveté|noun|E0604057 • Nonnaivete (in LEXICON)
trademark=	1	**E0522540**: trademark=Bacillus Calmette-Guérin (BCG), substrain Connaught • Bacillus Calmette-Guerin (BCG), substrain Connaught
Total non-ASCII line	4,345 (0.15 %) ⇨ 392 not-lex (0.01%)	
Total line in LEXICON	2,804,920	
Total records with non-ASCII	1,673 (0.39 %) ⇨ 62 not-Lex (0.01%)	
Total records in LEXICON	432,822	

II. LEXICON, Lexical Tools, & MetaMap

Lexical Tools use LEXICON and to generate database tables for mutation in various flow components. The SPECIALIST LEXICON and Lexical Tools have been upgraded its IO from pure ASCII to Unicode (UTF-8) since 2004. ‘C’ codes perform some functions as Lexical tools (referred as ‘C’ codes) and was used in MetaMap. ‘C’ codes use ASCII LEXICON for its inflectional, derivational, and other mutations. There are slightly difference from the results of Java Lexical Tools and ‘C’ codes because 1) The difference of ASCII LEXICON and LEXICON 2). The deficiency in algorithm implementation (please refer to reports on inflectional morphology and derivational morphology). This report will focus on the item 1) to discuss issues caused by ASCII conversion. The diagram below shows the framework of the relationship between LEXICON, Lexical Tools, and MetaMap.



III. Issues with Examples and Proposed Solutions

There are several issues raised by the ASCII conversion to cause different results between Java Lexicon tools and ‘C’ codes in the above framework and are described in details below:

Issue- 1. The ASCII conversion creates words are not known to LEXICON

This is the main problem and causes many differences in retrieving citation forms, uninflected forms, inflectional variants, spelling variants, etc. for records with Unicode characters. They can be further categorized as follows:

- *ASCII converted citation is not known to LEXICON (26)*
  - ⇨ No spelling variant exist
    1. 'Candomble', E0563996
    2. 'Koshland-Nemethy-Filmer model', E0524522
    3. 'Labbe vein', E0556296
    4. 'Labbe\'s vein', E0556297
    5. 'Loffler\'s alkaline methylene blue stain', E0527716
    6. 'Muthing', E0573093
    7. 'Nuevo Leon', E0683032
    8. 'Republique Federale Islamique des Comores', E0558494
    9. 'Ruther', E0571293
    10. 'Santo Andre', E0661148
    11. 'Schutz\'s fasciculus', E0527588
    12. 'Thormahlen\'s test', E0634501
    13. 'Vannas-Tubingen spring scissors', E0530928
    14. 'a trois', E0524660
    15. 'deja raconte', E0547066
    16. 'folie a trois', E0524662
    17. 'menage a trois', E0524661
    18. 'secondary Sjogren\'s syndrome', E0683091
    19. styrenation, E0586236
    20. 'tache cerebrale', E0547094

  - ⇨ Has spelling variants, but none of them are pure ASCII
    1. 'Munoz-Gonzalez', E0668220
    2. 'Pena-Quintana', E0641878
    3. 'Vazquez-Quintana', E0641879

  - ⇨ Has spelling variant, but none of them are equal to this form
    1. 'IFN-stimulated gene factor 3alpha', E0632531
    2. 'Spiculopteragia bohmi', E0534078
    3. 'interferon-stimulated gene factor 3alpha', E0632532

- *ASCII converted citation is same as other records (3)*
  - ⇨ E0543077|divorcé -> E0023635|divorce
  - ⇨ E0561275|Vitória -> E0561276|Vitoria
  - ⇨ E0571036|Böcking -> E0571034|Bocking

- ASCII converted spelling variant is not known to LEXICON:
  - E0669392| 5-stranded beta sheet
  - E0638200|Adelta fiber
  - E0690046| CTLA2-beta
  - E0504831|Cote d’Ivoire
  - E0670377|Delta psim
  - E0666928|
  - E0654835|
  - E0002583|
  - …

- ASCII converted abbreviations, acronyms, nominalization, etc. are not known to LEXICON

Example:

To get the citation of “Delta psiM”. The lexical record is:

```
{base=deltapsi(m)
spelling_variant=DeltaPsi(m)
spelling_variant=Deltapsi(m)
spelling_variant=DeltaPsim
spelling_variant=DeltaPsi m
spelling_variant=DeltaPsi M
spelling_variant=Delta Psi m
spelling_variant=Delta psi m
spelling_variant=Delta Psi M
spelling_variant=delta psi m
spelling_variant=deltapsi m
spelling_variant=Deltapsi m
spelling_variant=deltapsi(M)
spelling_variant=Deltapsi(M)
spelling_variant=delta psi(M)
spelling_variant=Delta Psi(M)
spelling_variant=DeltaPsi(M)
spelling_variant=Delta Psi(m)
spelling_variant=delta psi(m)
spelling_variant=ΔΨ(m)
spelling_variant=ΔΨ (m)
spelling_variant=ΔΨm
spelling_variant=ΔΨ m
spelling_variant=ΔΨ M
spelling_variant=ΔΨM
spelling_variant=ΔΨ(M)
spelling_variant=ΔΨ (M)
spelling_variant=Δ ψm
spelling_variant=Δψm
spelling_variant=Δ ψM
spelling_variant=ΔψM
spelling_variant=Δ ψ(M)
spelling_variant=Δψ(M)
spelling_variant=Δ ψ(m)
spelling_variant=Δψ(m)
spelling_variant=Deltapsim
entry=E0670377
    cat=noun
    variants=uncount
    abbreviation_of=mitochondrial membrane potential change|E0670369
    abbreviation_of=change in mitochondrial membrane potential
    abbreviation_of=change in mitochondrial transmembrane potential
}
```

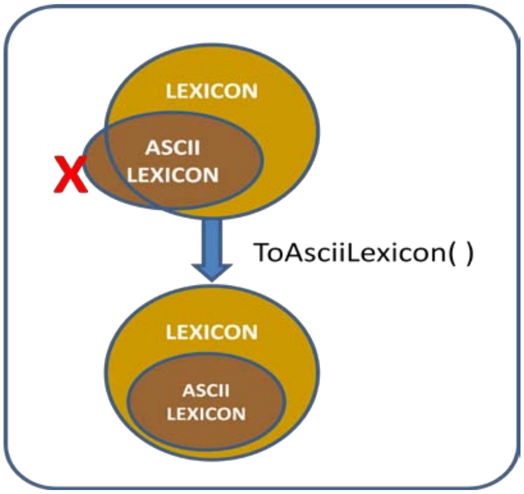
“spelling_variant=Δ ψM” is converted to “spelling_variant=Delta psiM” during the ASCII conversion. However, “Delta psiM” is not known to LEXICON (not a real word) and thus no citation form can be found.

Solution:

Theoretically, a lexical record should record everything related to the citation of the lexical record. Accordingly, there are only two possibilities when an ASCII converted word (such as spelling variant of the citation) is not known to LEXICON (does not exist in LEXICON):

- The ASCII converted word is not a real word
  - ⇨ Such words should be removed from ASCII LEXICON
- The Lexical record does not record everything
  - ⇨ Such words should be added to LEXICON

This problem can be resolved by developing an enhanced algorithm of the program, ToAsciiLexicon, to 1). Perform ASCII conversion 2). Detect and remove for ASCII conversions are not known to LEXICON 3). Report all found not known ASCII conversions for linguists to review and add/modified LEXICON if needed (please see Issue-4 for details). As discussed previously, ‘C’ codes use ASCII LEXICON to perform lexical mutations. The results won’t be correct if the input data (ASCII LEXICON) is not correct (include words not known to LEXICON). Please note that an ASCII Java Lexical Tools (based on current ASCII LEXICON) will results in same (wrong) results as ‘C’ codes for this category. Theoretically, the ASCII LEXICON should be a subset of the Unicode LEXICON (as shown in the diagram below) so that the results will be consistent between ‘C’ codes and Java Lexical Tools (since the results from ‘C’ codes should be a subset of the results from Java Lexical Tools). The proposed new version of ToAsciiLexicon program is to generate the correct ASCII LEXICON to resolve this issue correctly.



Issue- 2. The ASCII conversion creates wrong information/relationship in LEXICON

Problem:

The other problem is the ASCII conversion generates wrong information in the ASCII lexical records (LEIXCON) and causes ‘C’ codes to generate wrong information.

Example:

To get the citation of “mum”. This problem happens in the lexical record of “mu”:

```
{base=mu
spelling_variant=µ
spelling_variant=µm
entry=E0041164
    cat=noun
    variants=inv
    variants=metareg
    abbreviation_of=micrometer|E0040123
}
```

“spelling_variant=µm” is converted to “spelling_variant=mum“ and the ASCII lexical record becomes:

```
{base=mu
spelling_variant=mu
spelling_variant=mum
entry=E0041164
    cat=noun
    variants=inv
    variants=metareg
    abbreviation_of=micrometer|E0040123
}
```

The spelling variant “µm” is converted to “mum” and creates a relationship to the lexical record of “mum”. Thus, “C-code” generates “mu” (as well as “mum”) as citations of “mum” from the ASCII LEXICON. Other terms for this case include (E0647954|Galphai to E0651755|G alpha(i)) and (E0571034|Bocking to E0571036|Böcking), etc.. Please note that an ASCII Lexical Tools (based on current ASCII LEXICON) will result in the same mistakes and generate same result as ‘C’ codes. However, this is a wrong result. This problem does not exist through using Java Lexical Tools APIs in the java-Prolog interface.

Solution:

Implement intelligence in ToAsciiLexicon program to remove not-related term after ASCII conversion. In other word, “spelling_variant=mum” will be removed as shown in below:

```
{base=mu
spelling_variant=mu
entry=E0041164
    cat=noun
    variants=inv
    variants=metareg
    abbreviation_of=micrometer|E0040123
}
```

Issue-3. The results from Lexical Tools contain non-ASCII characters

Problem:

Lexical tools deals with UTF-8 (Unicode) and the results from its APIs is Unicode. Problems arise when the results from Lexical Tools contain non-ASCII characters.

Example:

To get the citation of “Aicardi-Goutieres syndrome”. The lexical record is:

```
{base=Aicardi-Goutières syndrome
spelling_variant=Aicardi-Goutieres syndrome
entry=E0572939
    cat=noun
    variants=uncount
}
```

Its ASCII lexical record is shown as below by two operations in the current ToAsciiLexicon:

1). “base=Aicardi-Goutières syndrome “ is converted to ““base=Aicardi-Goutieres syndrome “

2). “spelling_variant=Aicardi-Goutieres syndrome” is removed since it is the same as the converted citation form.

```
{base=Aicardi-Goutieres syndrome
entry=E0572939
    cat=noun
    variants=uncount
}
```

The result from ‘C’ code is to return the citation form (“Aicardi-Goutieres syndrome”) of this record. However, the citation form from Lexical Tools is: “Aicardi-Goutières syndrome”, which contains non-ASCII characters and is different from the result from ‘C” to cause problem.

Solution:

Implement the following algorithm in the “java-Prolog Interface”:

1). Apply ToAscii (from Lexical Tools):

The found citation form,“Aicardi-Goutières syndrome”, is converted to “Aicardi-Goutieres syndrome”

2). Verify if the ASCII conversion, “Aicardi-Goutieres syndrome”, is known to LEXICON:

- ⇨ If so, sent the result to MetaMap.
- ⇨ If not, remove it (return nothing).

Issue-4. Other related non-ASCII Issues and proposed solutions

Problem:

As discussed in Issue-1, it is possible that the SPECIALIST LEXICON does not include everything for the term in its lexical record (nothing is not perfect). We found out ten lexical records contain non-ASCII characters might have potential errors. They are:

- Records should be combined:
    - E0632882|PDGFR-alpha and E0638286|PDGFRA
    - E0644363|TCRbeta and E0680947|TCRB
    - E0693785|beta B1-crystallin and E0651383|betaB1-crystallin

    - E0543077|divorcé and E0023635|divorce
    - E0561275|Vitória and E0561276|Vitoria
    - E0571036|Böcking and E0571034|Bocking
- Issue of ö (convert to oe, not o)
    - E0532023| Koehler's first disease
    - E0585634| roentgenization
    - E0585635| roentgenize
    - E0683091| secondary Sjögren's syndrome
- Typo:
    - E0237602|Hand-Schuller-Christian syndrome
        - ⇨ spelling_variant=Hand Schuller Christian sydrome

Example 1:

```
{base=beta B1-crystallin
spelling_variant=beta B1 crystallin
spelling_variant=β B1-crystallin
spelling_variant=β-B1 crystallin
spelling_variant=β B1 crystallin
spelling_variant=βB1-crystallin
entry=E0693785
        cat=noun
        variants=reg
        variants=uncount
}
```

And

```
{base=betaB1-crystallin
spelling_variant=betaB1 crystallin
spelling_variant=beta B1 crystallin
spelling_variant=βB1-crystallin
spelling_variant=β B1-crystallin
spelling_variant=βB1 crystallin
spelling_variant=β B1 crystallin
entry=E0651383
        cat=noun
        variants=reg
        variants=uncount
}
```

Example 2:

```
{base=Hand-Schuller-Christian syndrome
spelling_variant=Hand Schuller Christian sydrome
spelling_variant=Hand-Schüller-Christian syndrome
spelling_variant=Hand Schüller Christian syndrome
spelling_variant=Hand-Schueller-Christian syndrome
entry=E0237602
      cat=noun
      variants=uncount
```

```
        variants=reg
    }
```

Solution:

Verify and correct LEXICON to make it right. As discussed in Issue-1, all non-known ASCII conversion terms will be reported for linguists to review and modify through LexBuild. Two plans for this case:

- For the existing incorrect records: will be fixed in 2012 LEXICON
- For the future (after LEXICON.2011) incorrect records: will be reviewed and corrected before annual release

IV. Conclusion

In conclusion, Lexical Systems Group will:

1) Implement a new enhanced algorithm on ToAsciiLexicon program and to:
   - Perform ASCII conversion
   - Detect and remove for ASCII conversions are not known to LEXICON
   - Report all found ASCII conversions that are not in LEXICON for linguists to review
   - Generate and deliver a new ASCII LEXICON of 2010 (ASAP) to MetaMap

ASCII issues of difference should be resolved with the new ASCII LEXICON.

2) Add new procedures to LEXICON annual generation to
   - Report all found ASCII conversions that are not in LEXICON for linguists to review
   - Correct found errors (from above) in LEXICON through LexBuild from 2012 release.

With this enhanced features, LEXICON could correct some errors and become better.